COMP9417: Machine Learning in the Unknown

April 11, 2021

Introduction

As a Data Scientist at Predictive Solutions Inc., you have become comfortable working with any type of dataset that comes your way. Increasingly, you have noticed that clients are becoming more and more concerned about security and privacy. They want to be able to use your machine learning knowledge without giving away any details about the data itself. For example, in a medical setting, hospitals would prefer not to disclose any private information about their patients nor the procedures they undergo. In such cases, the analyst has no *domain knowledge* to guide their modelling, and must instead rely solely on the algorithms and their training. Luckily, you have been trained well in COMP9417, so when your boss emails you:

`hey, new client is v. secretive, wants us to build a model for multi-class classification (see train.csv) - will only hire us long-term if our model can do better than their baseline on test.csv. let me know when you have a model up and running.`

You barely think about it before responding:

`No problem, will be ready late april.`

You then open up the files to notice that your boss was not kidding, there is absolutely no context for the data whatsoever. You quickly message the most reliable (3 to 5) analysts in your team, and they agree to help you out.

The actual data will be released on April 5, 2021.

Dataset & Evaluation

The dataset is composed of a `class` column, representing the target variable and which takes one of 6 possible values, and the remaining columns are 128 real valued features. You will use this training data to construct a classifier which you will apply to a provided test set. The test set contains only values of the features. You will need to submit your best predictions of the test set classes which will be used to evaluate your final model. Predictions will be evaluated using a class-weighted f1 score. The link to submit your predictions will be added here when the data is released. Update: we have provided you with the following in csv format: X_train, X_val, X_test, y_train, y_val it is completely up to you and your group to determine how best to make use of this. Recall that you will need to use best practices to come up with a model that generates predictions for X_test which will be submitted for evaluation and will count towards your final grade.

Update 2: You must submit a `y_test.csv` file that has a single column without a header (exactly the same format as `y_train.csv` containing your predictions. The order of these predictions must match the order in `X_test.csv`. Your code submission can either be a .py or .ipynb, or a .zip file with multiple .py/.ipynb component files.

Objectives

In this project, your group will use what they have learned in COMP9417 to construct a multi-class classifier as well as write a detailed report outlining your exploration of the data and approach to modelling.